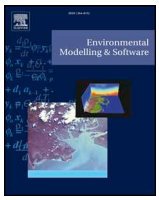
Environmental Modelling and Software 110 (2018) 1–2

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft



Preface: Environmental Data Science and Decision Support: Applications in Climate Change and the Ecological Footprint

Data Science is an emergent field that tries to take advantage of the current availability of data to bridge the gap between data and decisions. The borders of the concept itself have been controversial and in a previous and recent special issue (Environmental Modelling and Software, vol 106) Gibert et al. (2018) examine the scope of the Data Science concept and provide a modern view and a new definition: "*We consider Data Science as the multidisciplinary field that combines data analysis with data processing methods and domain expertise, transforming data into understandable and actionable knowledge relevant for informed decision making*". Data Science involves intensive consumption of available data, and does not necessarily imply only "big data" but also complexity. Because of the strategic added value that Data Science can provide to organizations and institutions, Data Science continues to increase in interest in all scientific and industrial sectors, including Environmental Sciences. The Gibert et al., (2018) work, analyzes in detail the new skills required to become a good data scientists as well as some possibilities of application in the environmental field. The paper ends raising current challenges in the area of Environmental Data Science and identifying new directions for contributions in the near future in this field. After the first paper, volume 106 contains eight other manuscripts, regarding applications in the fields of air quality and water cycle (from rainfall to water distribution networks and wastewater treatment plants), showing a variety of situations in which different data science processes contribute to environmental issues.

While volume 106 was primarily focusing of natural environmental systems, the current issue focuses on one of the most relevant challenges in environmental sciences: the consequences of human activity on the environment. This issue provides a sample of how Data Science can be used to better understand climate change and ecological footprint with ten papers related to planetary heating, climate modelling, and air, soil and water quality in both urban and non-urban areas, including wildlife in burnt areas, energy generation, including renewable energies and industrial symbiosis.

According to Cambridge Dictionary "ecological footprint'" is "*the effect that a person, company, activity, etc. has on the environment"* and it explicitly declares that *"Every organization should work towards a zero environmental footprint by conserving, restoring, and replacing the natural resources used in its operations".* In short, ecological footprint measures human impact on Earth's ecosystem and reveals the dependence of the human economy on natural capital. In fact, human activity has shown to be often aggressive and non-respectful of nature, and the environment is deteriorating at higher rate than is desirable. Indeed, predictions in 2000 expected that by 2013, humanity would be using natural capital 1.6 times as fast as nature can renew it [Wachernagek et al, 2000]. This has dramatic consequences on sustainability [Chambers et al, 2000] and there is urgent need to better understand where the main risks are and how we can reduce the ecological footprint to guarantee a longer life for the whole planet and, in the last term, for our own civilization.

This thematic issue of Environmental Modelling & Software reaches a bit further than strict Data Science applications and provides some ideas about how data-intensive processes from Environmental Data Science can feed on-line or off-line intelligent decision support systems which provide effective support to complex decision making in a variety of environmental scenarios. In the issue we can see different works on complete decision support systems and different architectures and uses, as well as their relationship with data science processes. Several works in the issue focus on sensor data and spatio-temporal modelling, but small data is also present. An interesting characteristic of the issue is the number of works facing hybrid environmental data science approaches where both knowledge-based and data-driven methods are combined to extract patterns from complex environmental systems. The work presented in this issue provides both methods for using data to better understand how human activity impacts on the environment, and better information to guide more sustainable activities as well, which in our opinion is one of the most relevant values of the Environmental Data Science.

Also, as one of the most critical issues in Data Science is to determine the proper Data Mining method to be used to analyze data, and as one of the challenges raised in Gibert et al. (2018) is the lack of guidelines in this respect, the issue begins with a methodological paper entitled *Which method to Use? An assessment of Data Mining Methods in Environmental Data Science,* which provides a global conceptual map for data mining methods (DMMCM) containing most popular methods used in environmental sciences, and a new methodology to use the DMMCM map for determining a suitable data mining method for a real environmental problem. An important characteristic of the proposal is that both the main environmental goal and the nature of the available data are considered to determine the data mining method, rather than the more classical approach of adapting data by hard transformations to fit the specifications of methods at the cost of losing interpretability.

The paper *Creating Weather Time Series through a Quantile Regression Ensemble* is an international collaboration between UK, Spain and Brazil that explores the use of quantile regression to provide best representatives of weather parameters (like temperature) to build time series that better estimate heat waves. The work advances regarding the traditional use of max-min daily values in these kind of sensor-data series. Weather series are essential for public health reasons, and for building energy consumption, building overheating, and occupants' thermal comfort as well. Given that every year more frequent and

https://doi.org/10.1016/j.envsoft.2018.11.006

severe heatwave events occur, improving the estimation of weather series has benefits in environmental management.

The rapid rate of climate warming in Alaska has raised interest for decision support tools in that area. However, available software do not provide sufficiently precise information at a regional level. The paper *Downscaling of climate model output for Alaskan stakeholders* responds to these needs by providing a downscaled climate model available in Alaska with a focus on better local predictions and visualization.

Continuing with the climate theme, the paper, *A novel approach to forecast urban surface-level ozone considering heterogeneous locations and limited information*, written in collaboration between the European Commission and Spain, shows how Data Science approaches can be used to forecast ozone in urban environments to advise citizens on unhealthy exposures of this air pollutant. A novel forecasting method for scarce information about ozone precursors is presented and compared with a pattern sequence based algorithm for spatiotemporal missing data imputation.

The paper *Inverse modelling for snow depths* improves prediction of snow depth in sub-alpine regions. The paper is linked to decision support software providing snow depth based on hourly simulation of the energy balance, the water equivalent and melt rates of snow cover. The inputs required for the simulation (emissivity, density of snow, etc.) are spatio-temporally conditioned on location and season. The paper proposes simultaneous Bayesian estimation of the vector of input parameters, based on routinely observed operational snow data at a local point. This is an interesting approach for estimation of several non-independent parameters, taking advantage of the large availability of snow data from monitoring systems

The paper *A knowledge modelling framework for intelligent environmental decision support systems and its application to some environmental problems* gives an holistic view of a general purpose decision support system architecture, which incorporates both statistical and artificial intelligence tools in a data science approach to support complex environmental decisions. The system includes ontologies, several data mining methods, and Bayesian networks for integrated environmental modelling. Data science processes are used to generate a knowledge base for the decision support part, which integrates domain knowledge provided by humans and literature, knowledge inducted with data mining methods and probabilistic knowledge inducted from Bayesian networks. Uncertainty is managed as well. The work shows applications in water resource management and surface water pollution, air pollution analysis and air pollution short term prediction, and soil pollution.

The paper *Adsorption characteristics of layered soil as delay barrier 2 of some organic contaminants: experimental and 3 numerical modeling* compares data-driven models with mechanistic models and simulation in the context of soil contamination modelling and the case study regards soils from an industrial zone in Tunisia. In fact, human activity is a major cause of soil contamination. Here, small samples and simulation are combined, as measurements are expensive.

The paper *Studying the Occurrence and Burnt Area of Wildfires using Zero-One-Inflated Structured Additive Beta Regression* proposes structured additive regression models and beta Distribution for studying wildfire occurrence and burnt area simultaneously. The case study focuses on an extraordinary situation in Galicia, Spain, where 2060 wildfires during the first half of August 2006. Fires are also a consequence of human activity in most of the cases, with enormous consequences on wildlife. Here, we can see another application related with characteristics of soil rather different from the previous ones.

The paper *Effects of the Pre-processing Algorithms in Fault Diagnosis of Wind Turbines* uses operational data, as with other papers in this issue. However, we can see here another use of this kind of available data. Machine learning techniques are used over SCADA data from wind turbines to learn about operation and maintenance services required (responsible for a significant portion of power generation cost.) The work evaluates the impact of systematically removing outliers on this type of sensor data and shows that, whereas this removal provides better accuracy rates in the training step, performs worse in the real-time testing, as important information of the system malfunction is missed when outliers are automatically removed.

The paper *A semantic multi-criteria approach to evaluate different types of energy generation tecnologies* analyses the problem of the selection of the best electricity generation plant taking into account the characteristics of a certain region. The paper presents a new method that is able to handle both semantic data and numerical indicators. The method exploits information from historical data, not necessarily big data, conveniently preprocessed to obtain risk factor measures, as well as the knowledge available in ontologies about waste byproducts and environmental pollutants. This decision support methods enables the automatic treatment of large sets of qualitative data, avoiding the biased results that would be obtained when these kind of indicators were excluded.

Finally, the paper *The influence of knowledge in the design of a recommender system to facilitate industrial symbiosis Markets* is related to a practice that can contribute to reducing ecological footprint: Industrial symbiosis. In this work, a recommender of promising industrial symbioses is proposed, by analyzing the role of specific domain knowledge in the performace of the system. Facilitating this kind of synergy, byproducts can be recycled as a second source of products in other enterprises and this contributes to increased sustainability.

The selection of papers for this issue was done with a rigorous blind peer-review process and high rate of rejection. From all the papers received, the papers most favorably evaluated by reviewers were selected. The reviewing process involved more than 50 international reviewers with the highest standards of expertise. The selected papers provide a nice overview of research conducted in multiple countries, including international collaborations and some collaborations between academia and corporations. With this issue, our purpose is to provide a useful perspective of how Data Science can contribute to reduce ecological footprint and increase sustainability, by disclosing at the same time the most critical parts of the Data Science process, which is to select a proper data mining method for the analysis.

The Guest Editors wish to thank the authors for contributing their original papers and for their patience throughout the peer review and publication process. We would also like to express gratitude to the numerous reviewers who contributed in constructively reviewing and improving the quality of the contributions within this issue.

Reference

Gibert, Karina, Horsburgh, J., Athanasiadis, I., Holmes, G., 2018. "Environmental Data Science.". Environmental Modelling & Software 106, 4–12. https://doi.org/10.1016/j.envsoft.2018.04.005.

Editors
Karina Gibert ((fiEMSs))
Dep. Statistics and Operations Research, Knowledge Engineering and Machine Learning group at Intelligent Data Science and Artificial Intelligence Research Center, Research Institute on Science and Technology for Sustainability, Universitat Politècnica de Catalunya-BarcelonaTech, Spain